
Tech Report

A Variational HEM Algorithm for Clustering Hidden Markov Models

Emanuele Coviello*

Department of Electrical and Computer Engineering
University of California, San Diego
9500 Gilman Drive La Jolla, CA 92093
ecoviell@ucsd.edu

Anotni B. Chan

Department of Computer Science
City University of Hong Kong
Tat Chee Avenue, Kowloon Tong, Hong Kong
abchan@cityu.edu.hk

Gert R.G. Lanckriet

Department of Electrical and Computer Engineering
University of California, San Diego
9500 Gilman Drive La Jolla, CA 92093
gert@ece.ucsd.edu

Abstract

The hidden Markov model (HMM) is a generative model that treats sequential data under the assumption that each observation is conditioned on the state of a discrete hidden variable that evolves in time as a Markov chain. In this paper, we derive a novel algorithm to cluster HMMs through their probability distributions. We propose a hierarchical EM algorithm that i) clusters a given collection of HMMs into groups of HMMs that are similar, in terms of the distributions they represent, and ii) characterizes each group by a “cluster center”, i.e., a novel HMM that is representative for the group. We present several empirical studies that illustrate the benefits of the proposed algorithm.

1 Introduction

The hidden Markov model (HMM) [1] is a probabilistic model that assumes a signal is generated by a double embedded stochastic process. A hidden state process, which evolves over discrete time instants as a Markov chain, encodes the dynamics of the signal, and an observation process, at each time conditioned on the current state, encodes the appearance of the signal. HMMs have been successfully applied to a variety of applications, including speech recognition [1], music analysis [2], on-line hand-writing recognition [3], analysis of biological sequences [4].

The focus of this paper is an algorithm for clustering HMMs. More precisely, we design an algorithm that, given a collection of HMMs, partitions them into K clusters of “similar” HMMs, while also learning a representative HMM “cluster center” that summarizes each cluster appropriately. This is

*<http://acsweb.ucsd.edu/~ecoviell/>

similar to standard k-means clustering, except that the data points are HMMs now instead of vectors in \mathbb{R}^d . Such HMM clustering algorithm has various potential applications, ranging from hierarchical clustering of sequential data (e.g., speech or motion sequences), over hierarchical indexing for fast retrieval, to reducing the computational complexity of estimating mixtures of HMMs from large datasets via hierarchical modeling (e.g., to learn semantic annotation models for music and video).

One possible approach is to group HMMs in *parameter* space. However, as HMM parameters lie on a non-linear manifold, they cannot be clustered by a simple application of the k-means algorithm, which assumes real vectors in a Euclidean space. One solution, proposed in [5], first constructs an appropriate similarity matrix between all HMMs that are to be clustered (e.g., based on the Bhattacharya affinity, which depends on the HMM parameters [6]), and then applies spectral clustering. While this approach has proven successful to group HMMs into similar clusters [5], it does not allow to generate novel HMMs as cluster centers. Instead, one is limited to representing each cluster by one of the given HMMs, e.g., the HMM which the spectral clustering procedure maps the closest to each spectral clustering center. This may be suboptimal for various applications of HMM clustering.

An alternative to clustering the HMMs in parameter space is to cluster them *directly* with respect to the *probability distributions* they represent. To cluster *Gaussian* probability distributions, Vasconcelos and Lipmann [7] proposed a hierarchical expectation-maximization (HEM) algorithm. This algorithm starts from a Gaussian mixture model (GMM) with $K^{(b)}$ components and reduces it to another GMM with fewer components, where each of the mixture components of the reduced GMM represents, i.e., *clusters*, a group of the original Gaussian mixture components. More recently, Chan et al. [8] derived an HEM algorithm to cluster *dynamic texture* (DT) models (i.e., linear dynamical systems, LDSs) through their probability distributions. HEM has been applied successfully to construct GMM hierarchies for efficient image indexing [9], to cluster video represented by DTs [10], and to estimate GMMs or DT mixtures (DTMs, i.e., LDS mixtures) from large datasets for semantic annotation of images [11], video [10] and music [12, 13].

To extend the HEM framework for GMMs to mixtures of HMMs (H3Ms), additional marginalization of the hidden-state processes is required, as for DTMs. However, while Gaussians and DTs allow tractable inference in the E-step of HEM, this is no longer the case for HMMs. Therefore, in this work, we derive a variational formulation of the HEM algorithm (VHEM) to cluster HMMs through their probability distributions, based on a variational approximation derived by Hershey [14]. The resulting algorithm not only allows to cluster HMMs, it also learns novel HMMs that are representative centers of each cluster, in a way that is consistent with the underlying generative probabilistic model of the HMM. The resulting VHEM algorithm can be generalized to handle other classes of graphical models, for which standard HEM would otherwise be intractable, by leveraging similar variational approximations. The efficacy of the VHEM-H3M algorithm is demonstrated for various applications, including hierarchical motion clustering, semantic music annotation, and online hand-writing recognition.

2 The hidden Markov model

A hidden Markov model (HMM) \mathcal{M} assumes a sequence of τ observations $y_{1:\tau}$ is generated by a double embedded stochastic process, where each observation y_t at time t depends on the state of a discrete hidden variable x_t and where the sequence of hidden states $x_{1:\tau}$ evolves as a first order Markov process. The discrete variables can take one of N values, and the evolution of the hidden process is encoded in a transition matrix $A = [a_{\beta,\gamma}]_{\beta,\gamma=1,\dots,N}$ whose entries are the state transition probabilities $a_{\beta,\gamma} = P(x_{t+1} = \gamma | x_t = \beta)$, and an initial state distribution $\pi = [\pi_1, \dots, \pi_N]$, where $\pi_\beta = P(x_1 = \beta)$. Each state generates observation accordingly to an emission probability density function, $p(y_t | x_t = \beta, \mathcal{M})$, which here we assume to be a Gaussian mixture model:

$$p(y|x = \beta) = \sum_{m=1}^M c_{\beta,m} \mathcal{N}(y; \mu_{\beta,m}, \Sigma_{\beta,m}) \quad (1)$$

where M is the number of Gaussian components and $c_{\beta,m}$ are the mixing weights. In the following, when referring to a sequences of length τ , we will use the notation $\pi_{x_{1:\tau}} P(x_{1:\tau}) = \pi_{x_1} \prod_{t=2}^{\tau} a_{x_{t-1}, x_t}$ to represent the probability that the HMM generates the state sequence $x_{1:\tau}$. The HMM is specified by the parameters $\mathcal{M} = \{\pi, A, c_{\beta,m}, \mu_{\beta,m}, \Sigma_{\beta,m}\}$ which can be efficiently learned with the forward-backward algorithm [1], which is based on maximum likelihood.

A hidden Markov mixture model (H3M) [15] models a set observation sequences as samples from a group of K hidden Markov models, which represent different sub-behaviors. For a given sequence, an assignment variable $z \sim \text{multinomial}(\omega_1, \dots, \omega_K)$ selects the parameters of one of the K HMMs, where the k -th HMM is selected with probability ω_k . Each mixture component is parametrized by $\mathcal{M}_z = \{\pi^z, A^z, c_{\beta,m}^z, \mu_{\beta,m}^z, \Sigma_{\beta,m}^z\}$ and the H3M is parametrized by $\mathcal{M} = \{\omega_z, \mathcal{M}_z\}_{z=1}^K$. Given a collection $\mathcal{S} = \{y_{1:\tau}^1, \dots, y_{1:\tau}^{|\mathcal{S}|}\}$ of relevant observation sequences, the parameters of \mathcal{M} can be learned with recourse to the EM algorithm [15].

To reduce clutter, here we assume that all the HMMs have the same number of states N and that all emission probabilities have M mixture components, though a more general case could be derived.

3 Variational hierarchical EM algorithm for H3Ms

The hierarchical expectation maximization algorithm (HEM) [7] was initially proposed to cluster Gaussian distributions, by reducing a GMM with a large number of components to a new GMM with fewer components, and then extended to dynamic texture models [8]. In this section we derive a variational formulation of the HEM algorithm (VHEM) to cluster HMMs.

3.1 Formulation

Let $\mathcal{M}^{(b)}$ be a base hidden Markov mixture model with $K^{(b)}$ components. The goal of the VHEM algorithm is to find a reduced mixture $\mathcal{M}^{(r)}$ with $K^{(r)} < K^{(b)}$ components that represent $\mathcal{M}^{(b)}$. The likelihood of a random sequence $y_{1:\tau} \sim \mathcal{M}^{(b)}$ is given by

$$p(y_{1:\tau}|\mathcal{M}^{(b)}) = \sum_{i=1}^{K^{(b)}} \omega_i^{(b)} p(y_{1:\tau}|z^{(b)} = i, \mathcal{M}^{(b)}), \quad (2)$$

where $z^{(b)} \sim \text{multinomial}(\omega_1^{(b)}, \dots, \omega_{K^{(b)}}^{(b)})$ is the hidden variable that indexes the mixture components. $p(y_{1:\tau}|z = i, \mathcal{M}^{(b)})$ is the likelihood of $y_{1:\tau}$ under the i th mixture component, and $\omega_i^{(b)}$ is the prior weight for the i th component. The likelihood of the random sequence $y_{1:\tau} \sim \mathcal{M}^{(r)}$ is

$$p(y_{1:\tau}|\mathcal{M}^{(r)}) = \sum_{j=1}^{K^{(r)}} \omega_j^{(r)} p(y_{1:\tau}|z^{(r)} = j, \mathcal{M}^{(r)}), \quad (3)$$

where $z^{(r)} \sim \text{multinomial}(\omega_1^{(r)}, \dots, \omega_{K^{(r)}}^{(r)})$ is the hidden variable for indexing components in $\mathcal{M}^{(r)}$. Note that we will always use i and j to index the components of the base model, $\mathcal{M}^{(b)}$, and the reduced model, $\mathcal{M}^{(r)}$, respectively. In addition, we will always use β and γ to index the hidden states of $\mathcal{M}_i^{(b)}$ and ρ and σ for $\mathcal{M}_j^{(r)}$. To reduce clutter we will denote $p(y_{1:\tau}|z^{(b)} = i, \mathcal{M}^{(b)}) = p(y_{1:\tau}|\mathcal{M}_i^{(b)})$, and $\mathbb{E}_{y_{1:\tau}|\mathcal{M}^{(b)}, z^{(b)}=i}[\cdot] = \mathbb{E}_{\mathcal{M}_i^{(b)}}[\cdot]$. In addition, we will use short-hands $\mathcal{M}_{i,\beta_{1:\tau}}^{(b)}$ and $\mathcal{M}_{i,\rho_{1:\tau}}^{(r)}$ when conditioning over specific state sequences. For example, we denote $p(y_{1:\tau}|x_{1:\tau} = \beta_{1:\tau}, \mathcal{M}_i^{(b)}) = p(y_{1:\tau}|\mathcal{M}_{i,\beta_{1:\tau}}^{(b)})$, and $\mathbb{E}_{y_{1:\tau}|\mathcal{M}_i^{(b)}, x_{1:\tau}=\beta_{1:\tau}}[\cdot] = \mathbb{E}_{\mathcal{M}_{i,\beta_{1:\tau}}^{(b)}}[\cdot]$. Finally, we will use m and ℓ for indexing the gaussian mixture components of the emission probabilities of the base respectively reduced mixture, which we will denote as $\mathcal{M}_{\beta,m}^{(b),i}$ and $\mathcal{M}_{\rho,\ell}^{(r),j}$.

3.2 Parameter estimation - a variational formulation

To obtain the reduced model, we consider a set of N virtual samples drawn from the base model $\mathcal{M}^{(b)}$, such that $N_i = N\omega_i^{(b)}$ samples are drawn from the i th component. We denote the set of N_i virtual samples for the i th component as $Y_i = \{y_{1:\tau}^{(i,m)}\}_{m=1}^{N_i}$, where $y_{1:\tau}^{(i,m)} \sim \mathcal{M}_i^{(b)}$, and the entire set of N samples as $Y = \{Y_i\}_{i=1}^{K^{(b)}}$. Note that, in this formulation, we are not generating virtual samples $\{x_{1:\tau}^{(i,m)}, y_{1:\tau}^{(i,m)}\}$ according to the joint distribution of the base component, $p(x_{1:\tau}, y_{1:\tau}|\mathcal{M}_i^{(b)})$. The reason is that the hidden state spaces of each base mixture component $\mathcal{M}_i^{(b)}$ may have a different

representation, (e.g., the numbering of the hidden states may be permuted between the components). This basis mismatch will cause problems when the parameters of $\mathcal{M}_j^{(r)}$ are computed from virtual samples of the hidden states of $\{\mathcal{M}_i^{(b)}\}$. Instead, we must treat $X_i = \{x_{1:\tau}^{(i,m)}\}$ as “missing” information, as in the standard EM formulation.

The likelihood of the virtual samples is

$$\mathcal{J}(\mathcal{M}^{(r)}) = \log p(Y|\mathcal{M}^{(r)}) = \sum_{i=1}^{K^{(b)}} \log p(Y_i|\mathcal{M}^{(r)}) = \sum_{i=1}^{K^{(b)}} \log \sum_{j=1}^{K^{(r)}} \omega_j^{(r)} p(Y_i|\mathcal{M}_j^{(r)}). \quad (4)$$

In particular, for a given $\mathcal{M}^{(r)}$, the computation of $\log p(Y_i|\mathcal{M}^{(r)})$ can be carried out solving the optimization problems [16, 17]:

$$\log p(Y_i|\mathcal{M}^{(r)}) = \max_{\mathcal{P}_i(z_i)} \log p(Y_i|\mathcal{M}^{(r)}) - D(\mathcal{P}_i(z_i)||P(z_i = j|Y_i, \mathcal{M}^{(r)})) \quad (5)$$

$$= \max_{\mathcal{P}_i(z_i)} \sum_j \mathcal{P}_i(z_i = j) \left[\log \omega_j^{(r)} + \log p(Y_i|\mathcal{M}_j^{(r)}) - \log \mathcal{P}_i(z_i = j) \right] \quad (6)$$

for $i = 1, \dots, K^{(b)}$, where $\mathcal{P}_i(z_i)$ are variational distributions and $D(p||q) = \int p(y) \log \frac{p(y)}{q(y)} dy$ is the Kullback-Leibler (KL) divergence between two distributions, p and q . In order to obtain a consistent clustering [7], we assume the whole sample Y_i is assigned to the same component of the reduced model, i.e., $\mathcal{P}_i(z_i = j) = z_{ij}$, with $\sum_{j=1}^{K^{(r)}} z_{ij} = 1, \forall i$ and $z_{ij} \geq 0 \forall i, j$, and (6) becomes:

$$\log p(Y_i|\mathcal{M}^{(r)}) = \max_{z_{ij}} \sum_j z_{ij} \left[\log \omega_j^{(r)} + \log p(Y_i|\mathcal{M}_j^{(r)}) - \log z_{ij} \right] \quad (7)$$

Considering that virtual samples Y_i are independent for different values of i , we can solve (6) independently for each i , using the result in Section 6.2, and find

$$\hat{z}_{ij} = \frac{\omega_j^{(r)} \exp\{N_i \mathbb{E}_{\mathcal{M}_i^{(b)}} [\log p(Y_i|\mathcal{M}_j^{(r)})]\}}{\sum_{j'=1}^{K^{(r)}} \omega_{j'}^{(r)} \exp\{N_i \mathbb{E}_{\mathcal{M}_i^{(b)}} [\log p(Y_i|\mathcal{M}_{j'}^{(r)})]\}}. \quad (8)$$

For the likelihood of the virtual samples $p(Y_i|\mathcal{M}_j^{(r)})$ we use

$$\log p(Y_i|\mathcal{M}_j^{(r)}) = \sum_{m=1}^{N_i} \log p(y_{1:\tau}^{(i,m)}|\mathcal{M}_j^{(r)}) = N_i \left[\frac{1}{N_i} \sum_{m=1}^{N_i} \log p(y_{1:\tau}^{(i,m)}|\mathcal{M}_j^{(r)}) \right] \quad (9)$$

$$\approx N_i \mathbb{E}_{\mathcal{M}_i^{(b)}} [\log p(y_{1:\tau}|\mathcal{M}_j^{(r)})] \quad (10)$$

where (10) follows from the law of large numbers [7] (as $N_i \rightarrow \infty$). Substituting (10) in (8) we get the formula for z_{ij} derived in [7]:

$$\hat{z}_{ij} = \frac{\omega_j^{(r)} \exp\{N_i \mathbb{E}_{\mathcal{M}_i^{(b)}} [\log p(y_{1:\tau}|\mathcal{M}_j^{(r)})]\}}{\sum_{j'=1}^{K^{(r)}} \omega_{j'}^{(r)} \exp\{N_i \mathbb{E}_{\mathcal{M}_i^{(b)}} [\log p(y_{1:\tau}|\mathcal{M}_{j'}^{(r)})]\}}. \quad (11)$$

We follow a similar approach to compute the expected log-likelihoods $\mathbb{E}_{\mathcal{M}_i^{(b)}} [\log p(y_{1:\tau}|\mathcal{M}_j^{(r)})]$.

We introduce variational distributions $\mathcal{P}_{\beta_{1:\tau}}^{i,j}$ to approximate $P(x_{1:\tau}|y_{1:\tau}, \mathcal{M}_j^{(r)})$ for observations $y_{1:\tau} \sim \mathcal{M}_i^{(b)}$ emitted by state sequence $\beta_{1:\tau}$, and solve the maximization problem

$$\mathbb{E}_{\mathcal{M}_i^{(b)}} [\log p(y_{1:\tau}|\mathcal{M}_j^{(r)})] = \quad (12)$$

$$= \max_{\mathcal{P}^{i,j}} \sum_{\beta_{1:\tau}} \pi_{x_{1:\tau}}^{(b),i} \mathbb{E}_{\mathcal{M}_{i,\beta_{1:\tau}}^{(b)}} \left[\log p(y_{1:\tau}|\mathcal{M}_j^{(r)}) - D(\mathcal{P}_{\beta_{1:\tau}}^{i,j} || P(x_{1:\tau}|y_{1:\tau}, \mathcal{M}_j^{(r)})) \right] \quad (13)$$

$$= \max_{\mathcal{P}^{i,j}} \sum_{\beta_{1:\tau}} \pi_{x_{1:\tau}}^{(b),i} \mathbb{E}_{\mathcal{M}_{i,\beta_{1:\tau}}^{(b)}} \left[\sum_{\rho_{1:\tau}} \mathcal{P}_{\beta_{1:\tau}}^{i,j}(x_{1:\tau} = \rho_{1:\tau}) \log \frac{\pi_{\rho_{1:\tau}}^{(r),j} p(y_{1:\tau}|\mathcal{M}_{j,\rho_{1:\tau}}^{(r)})}{\mathcal{P}_{\beta_{1:\tau}}^{i,j}(x_{1:\tau} = \rho_{1:\tau})} \right] \quad (14)$$

where in (13) we have used the law of total probability to condition the expectation over each state sequence $\beta_{1:\tau}$ of $\mathcal{M}_i^{(b)}$

In general, maximizing (14) exactly sets the variational distribution to the true posterior and reduces (together with (11)) to the E-step of the HEM algorithm for hidden state models derived in [10]:

$$\mathbb{E}_{\mathcal{M}_i^{(b)}} [\log p(y_{1:\tau} | \mathcal{M}_j^{(r)})] = \mathbb{E}_{\mathcal{M}_i^{(b)}} [\mathbb{E}_{x_{1:\tau} | \hat{\mathcal{M}}_j^{(r)}} [\log p(x_{1:\tau}, y_{1:\tau} | \mathcal{M}_j^{(r)})]] + \tilde{\mathcal{H}} \quad (15)$$

where the inner expectation is taken with respect to the current estimate $\hat{\mathcal{M}}_j^{(r)}$ of $\mathcal{M}_j^{(r)}$, and $\tilde{\mathcal{H}}$ is some term that does not depend on $\mathcal{M}_j^{(r)}$.

3.3 The variational HEM for HMMs

The maximization of (14) cannot be carried out in a efficient way, as it involves computing the expected log-likelihood of a mixture. To make it tractable we follow a variational approximation proposed by Hershey [14], and restrict the maximization to factored distribution in the form of a Markov chain, i.e.,

$$\mathcal{P}_{\beta_{1:\tau}}^{i,j}(x_{1:\tau} = \rho_{1:\tau}) = \phi_{\rho_{1:\tau} | \beta_{1:\tau}}^{i,j} = \phi_1^{i,j}(\rho_1, \beta_1) \prod_{t=2}^{\tau} \phi_t^{i,j}(\rho_{t-1}, \rho_t, \beta_t) \quad (16)$$

where $\sum_{\rho=1}^N \phi_1^{i,j}(\rho_1, \beta_1) = 1 \forall \beta_1$ and $\sum_{\rho=1}^N \phi_t^{i,j}(\rho_{t-1}, \rho_t, \beta_t) = 1 \forall \beta_t, \rho_{t-1}$.

Substituting (16) into (14) we get a lower bound to the expected log-likelihood (14), i.e.,

$$\mathbb{E}_{\mathcal{M}_i^{(b)}} [\log p(y_{1:\tau} | \mathcal{M}_j^{(r)})] \geq \mathcal{J}^{i,j}(\mathcal{M}_j^{(r)}, \phi^{i,j}) \quad \forall \phi^{i,j} \quad (17)$$

where we have defined

$$\mathcal{J}^{i,j}(\mathcal{M}_j^{(r)}, \phi^{i,j}) = \sum_{\beta_{1:\tau}} \pi_{\beta_{1:\tau}}^{(b),i} \sum_{\rho_{1:\tau}} \phi_{\rho_{1:\tau} | \beta_{1:\tau}}^{i,j} \log \frac{\pi_{\rho_{1:\tau}}^{(r),j} \exp \mathbb{E}_{\mathcal{M}_{i,\beta_{1:\tau}}^{(b)}} [\log p(y_{1:\tau} | \mathcal{M}_{j,\rho_{1:\tau}}^{(r)})]}{\phi_{\rho_{1:\tau} | \beta_{1:\tau}}^{i,j}}. \quad (18)$$

Using the property of HMM with memory one that observations at different time instants are independent given the corresponding states, we can break the expectation term in equation (18) in the following summation

$$\mathbb{E}_{\mathcal{M}_{i,\beta_{1:\tau}}^{(b)}} [\log p(y_{1:\tau} | \mathcal{M}_{j,\rho_{1:\tau}}^{(r)})] = \sum_{t=1}^{\tau} L(\mathcal{M}_{i,\beta_t}^{(b)} || \mathcal{M}_{j,\rho_t}^{(r)}) \quad (19)$$

where $L(\mathcal{M}_{i,\beta_t}^{(b)} || \mathcal{M}_{j,\rho_t}^{(r)}) = \mathbb{E}_{\mathcal{M}_{i,\beta_t}^{(b)}} [\log p(y_t | \mathcal{M}_{j,\rho_t}^{(r)})]$. As the emission probabilities are GMMs, the computation (19) cannot be carried out efficiently. Hence, we use a variational approximation [18], and introduce variational parameters $\eta_{\ell|m}^{(i,\beta),(j,\rho)}$ for $\ell, m = 1, \dots, M$, with $\sum_{\ell=1}^M \eta_{\ell|m}^{(i,\beta),(j,\rho)} = 1 \forall m$, and $\eta_{\ell|m}^{(i,\beta),(j,\rho)} \geq 0 \forall \ell, m$. Intuitively, $\eta^{(i,\beta),(j,\rho)}$ is the responsibility matrix between gaussian observation components for state β in $\mathcal{M}_i^{(b)}$ and state ρ in $\mathcal{M}_j^{(r)}$, where $\eta_{\ell|m}^{(i,\beta),(j,\rho)}$ means the probability that an observation from component m of $\mathcal{M}_{i,\beta}^{(b)}$ corresponds to component ℓ of $\mathcal{M}_{j,\rho}^{(r)}$. Again, we obtain a lower bound:

$$L(\mathcal{M}_{i,\beta_t}^{(b)} || \mathcal{M}_{j,\rho_t}^{(r)}) \geq \mathcal{L}(\mathcal{M}_{i,\beta_t}^{(b)} || \mathcal{M}_{j,\rho_t}^{(r)}) \quad \forall \eta^{(i,\beta),(j,\rho)} \quad (20)$$

where we have defined:

$$\mathcal{L}(\mathcal{M}_{i,\beta_t}^{(b)} || \mathcal{M}_{j,\rho_t}^{(r)}) = \sum_{m=1}^M c_{\beta,m}^{(b),i} \sum_{\ell=1}^M \eta_{\ell|m}^{(i,\beta),(j,\rho)} \left[\log c_{\rho,\ell}^{(r),j} + L_G(\mathcal{M}_{\beta,m}^{(b),i} || \mathcal{M}_{\rho,\ell}^{(r),j}) - \log \eta_{\ell|m}^{(i,\beta),(j,\rho)} \right] \quad (21)$$

where $L_G(\mathcal{M}_{\beta,m}^{(b),i} || \mathcal{M}_{\rho,\ell}^{(r),j}) = \mathbb{E}_{y | \mathcal{M}_{\beta,m}^{(b),i}} [\log P(y | \mathcal{M}_{\rho,\ell}^{(r),j})]$ can be computed exactly for Gaussians

$$L_G(\mathcal{M}_{\beta,m}^{(b),i} || \mathcal{M}_{\rho,\ell}^{(r),j}) = -\frac{1}{2} d \log 2\pi + \log |\Sigma_{j,\rho}^{(r)}| - \frac{1}{2} \text{tr} \left(\Sigma_{j,\rho}^{(r)-1} \Sigma_{i,\beta}^{(b)} \right) \quad (22)$$

$$-\frac{1}{2} (\mu_{j,\rho}^{(r)} - \mu_{i,\beta}^{(b)})^T \Sigma_{j,\rho}^{(r)-1} (\mu_{j,\rho}^{(r)} - \mu_{i,\beta}^{(b)}). \quad (23)$$

Plugging (21) into (19) and (18) we get the lower bound to the expected log-likelihood:

$$\mathbb{E}_{\mathcal{M}_i^{(b)}} \left[\log p(y_{1:\tau} | \mathcal{M}_j^{(r)}) \right] \geq \mathcal{J}^{i,j}(\mathcal{M}^{(r)}, \phi^{i,j}, \eta) \quad \forall \phi^{i,j}, \eta \quad (24)$$

where we have defined:

$$\mathcal{J}^{i,j}(\mathcal{M}_j^{(r)}, \phi^{i,j}, \eta) = \sum_{\beta_{1:\tau}} \pi_{\beta_{1:\tau}}^{(b),i} \sum_{\rho_{1:\tau}} \phi_{\rho_{1:\tau} | \beta_{1:\tau}}^{i,j} [\log \pi_{\rho_{1:\tau}}^{(r),j} + \sum_{t=1}^{\tau} \mathcal{L}(\mathcal{M}_{i,\beta_t}^{(b)} || \mathcal{M}_{j,\rho_t}^{(r)}) - \log \phi_{\rho_{1:\tau} | \beta_{1:\tau}}^{i,j}]. \quad (25)$$

Maximizing the right hand side of (24) with respect to ϕ, η finds the most accurate approximation to the real posterior within the restricted class, i.e., the one that achieves the tightest lower bound.

Finally, plugging (24) into (4), we obtain a lower bound on the log-likelihood of the virtual sample:

$$\mathcal{J}(\mathcal{M}^{(r)}) \geq \mathcal{J}(\mathcal{M}^{(r)}, z, \phi, \eta) \quad \forall z, \phi, \eta \quad (26)$$

where we have defined:

$$\begin{aligned} \mathcal{J}(\mathcal{M}^{(r)}, z, \phi, \eta) &= \sum_{i=1}^{K^{(b)}} \sum_{j=1}^{K^{(r)}} z_{ij} \log \omega_j^{(r)} - \sum_{i=1}^{K^{(b)}} \sum_{j=1}^{K^{(r)}} z_{ij} \log z_{ij} \\ &+ \sum_{i=1}^{K^{(b)}} \sum_{j=1}^{K^{(r)}} z_{ij} N_i \sum_{\beta=1}^S \pi_{\beta_{1:\tau}}^{(b),i} \sum_{\rho=1}^S \phi_{\rho | \beta}^{i,j} \log \pi_{\rho_{1:\tau}}^{(r),j} \\ &+ \sum_{i=1}^{K^{(b)}} \sum_{j=1}^{K^{(r)}} z_{ij} N_i \sum_{\beta=1}^S \pi_{\beta_{1:\tau}}^{(b),i} \sum_{\rho=1}^S \phi_{\rho_{1:\tau} | \beta_{1:\tau}}^{i,j} \sum_{t=1}^{\tau} \mathcal{L}(\mathcal{M}_{i,\beta_t}^{(b)} || \mathcal{M}_{j,\rho_t}^{(r)}) \\ &- \sum_{i=1}^{K^{(b)}} \sum_{j=1}^{K^{(r)}} z_{ij} N_i \sum_{\beta=1}^S \pi_{\beta_{1:\tau}}^{(b),i} \sum_{\rho=1}^S \phi_{\rho_{1:\tau} | \beta_{1:\tau}}^{i,j} \log \phi_{\rho_{1:\tau} | \beta_{1:\tau}}^{i,j} \end{aligned} \quad (27)$$

To find the tightest possible lower bound to the log-likelihood of the virtual sample we need to solve

$$\mathcal{J}(\mathcal{M}^{(r)}) \geq \max_{z, \phi, \eta} \mathcal{J}(\mathcal{M}^{(r)}, z, \phi, \eta). \quad (28)$$

Starting from an initial guess for $\mathcal{M}^{(r)}$, the parameters can be estimated by maximizing (28) iteratively with respect to (*E-step*) η , ϕ , z and (*M-step*) $\mathcal{M}^{(r)}$

3.4 E-step

The E-steps first considers the maximization of (28) with respect to η for fixed $\mathcal{M}^{(r)}$, z and ϕ , i.e.,

$$\hat{\eta} = \operatorname{argmax}_{\eta} \mathcal{J}(\mathcal{M}^{(r)}, z, \phi, \eta) \quad (29)$$

It can be easily verified that the maximization (29) does not depend on z and ϕ can be carried out independently for each tuple (i, j, β, ρ, m) using the result in Section 6.2 [18], which gives:

$$\hat{\eta}_{\ell|m}^{(i,\beta),(j,\rho)} = \frac{c_{\rho,\ell}^{(r),j} \exp \left\{ L_G(\mathcal{M}_{\beta,m}^{(b),i} || \mathcal{M}_{\rho,\ell}^{(r),j}) \right\}}{\sum_{\ell'=1}^M c_{\rho,\ell'}^{(r),j} \exp \left\{ L_G(\mathcal{M}_{\beta,m}^{(b),i} || \mathcal{M}_{\rho',\ell'}^{(r),j}) \right\}} \quad (30)$$

and that the terms in (21) can then be computed for each (i, j, β, ρ) as:

$$L(\mathcal{M}_{i,\beta}^{(b)} || \mathcal{M}_{j,\rho}^{(r)}) = \sum_{m=1}^M c_{\beta,m}^{(b),i} \log \sum_{\ell=1}^M c_{\rho,\ell}^{(r),j} \exp \left\{ L_G(\mathcal{M}_{\beta,m}^{(b),i} || \mathcal{M}_{\rho,\ell}^{(r),j}) \right\}. \quad (31)$$

Next, (28) is maximized with respect to ϕ for fixed $(\mathcal{M}^{(r)}, z \text{ and } \eta)$, i.e.,

$$\hat{\phi} = \operatorname{argmax}_{\phi} \mathcal{J}(\mathcal{M}^{(r)}, z, \phi, \eta) \quad (32)$$

The maximization does not depends on \mathbf{z} and can be carried out independently for each pair (i, j) with a backward recursion [14] that computes

$$\hat{\phi}_t^{i,j}(\rho_{t-1}, \rho_t, \beta_t) = \frac{a_{\rho_{t-1}, \rho_t}^{(r),j} \exp \left\{ L(\mathcal{M}_{i,\beta_t}^{(b)} \parallel \mathcal{M}_{j,\rho_t}^{(r)}) + \mathcal{L}_{t+1}^{i,j}(\beta_t, \rho_t) \right\}}{\sum_{\rho} a_{\rho_{t-1}, \rho}^{(r),j} \exp \left\{ L(\mathcal{M}_{i,\beta_t}^{(b)} \parallel \mathcal{M}_{j,\rho}^{(r)}) + \mathcal{L}_{t+1}^{i,j}(\beta_t, \rho) \right\}} \quad (33)$$

$$\mathcal{L}_t^{i,j}(\rho_{t-1}, \beta_{t-1}) = \sum_{\beta=1}^N a_{\beta_{t-1}, \beta}^{(b),i} \log \sum_{\rho=1}^N a_{\rho_{t-1}, \rho}^{(r),j} \exp \left\{ L(\mathcal{M}_{i,\beta}^{(b)} \parallel \mathcal{M}_{j,\rho}^{(r)}) + \mathcal{L}_{t+1}^{i,j}(\beta, \rho) \right\} \quad (34)$$

for $T = \tau, \dots, 2$, where it is understood that $\mathcal{L}_{\tau+1}^{i,j}(\beta_t, \rho_t) = 0$, and terminates with

$$\hat{\phi}_1^{i,j}(\rho_1, \beta_1) = \frac{\pi_{\rho_1}^{(r),j} \exp \left\{ L(\mathcal{M}_{i,\beta_1}^{(b)} \parallel \mathcal{M}_{j,\rho_1}^{(r)}) + \mathcal{L}_2^{i,j}(\beta_1, \rho_1) \right\}}{\sum_{\rho} \pi_{\rho}^{(r),j} \exp \left\{ L(\mathcal{M}_{i,\beta_1}^{(b)} \parallel \mathcal{M}_{j,\rho}^{(r)}) + \mathcal{L}_2^{i,j}(\beta_1, \rho) \right\}} \quad (35)$$

$$\mathcal{J}^{i,j}(\mathcal{M}_j^{(r)}, \hat{\phi}^{i,j}, \boldsymbol{\eta}) = \sum_{\beta=1}^N \pi_{\beta}^{(b),i} \log \sum_{\rho=1}^N \pi_{\rho}^{(r),j} \exp \left\{ L(\mathcal{M}_{i,\beta}^{(b)} \parallel \mathcal{M}_{j,\rho}^{(r)}) + \mathcal{L}_2^{i,j}(\beta, \rho) \right\}. \quad (36)$$

Next, the maximization of (28) with respect to \mathbf{z} for fixed $\mathcal{M}^{(r)}$ ϕ and $\boldsymbol{\eta}$, i.e.,

$$\hat{\mathbf{z}} = \underset{\mathbf{z}}{\operatorname{argmax}} \mathcal{J}(\mathcal{M}^{(r)}, \mathbf{z}, \phi, \boldsymbol{\eta}) \quad (37)$$

reduces to compute the \hat{z}_{ij} as in (11) using (36) to approximate (10).

Finally, we compute the following summary statistics:

$$\nu_1^{i,j}(\sigma, \gamma) = \pi_{\gamma}^{(b),i} \hat{\phi}_1^{i,j}(\sigma, \gamma) \quad (38)$$

$$\xi_t^{i,j}(\rho, \sigma, \gamma) = \left(\sum_{\beta=1}^N \nu_{t-1}^{i,j}(\rho, \beta) a_{\beta, \gamma}^{(b),i} \right) \hat{\phi}_t^{i,j}(\rho, \sigma, \gamma) \text{ for } t = 2, \dots, \tau \quad (39)$$

$$\nu_t^{i,j}(\sigma, \gamma) = \sum_{\rho=1}^N \xi_t^{i,j}(\rho, \sigma, \gamma) \text{ for } t = 2, \dots, \tau \quad (40)$$

and the aggregates

$$\hat{\nu}_1^{i,j}(\sigma) = \sum_{\gamma=1}^N \nu_1^{i,j}(\sigma, \gamma) \quad (41)$$

$$\hat{\nu}^{i,j}(\sigma, \gamma) = \sum_{t=1}^{\tau} \nu_t^{i,j}(\sigma, \gamma) \quad (42)$$

$$\hat{\xi}^{i,j}(\rho, \sigma) = \sum_{t=2}^{\tau} \sum_{\gamma=1}^N \xi_t^{i,j}(\rho, \sigma, \gamma). \quad (43)$$

The quantity $\nu_t^{i,j}(\sigma, \gamma)$ is the responsibility between state γ of the HMM $\mathcal{M}_i^{(b)}$ and state σ of the HMM $\mathcal{M}_j^{(r)}$ at time t , when modeling a sequence generated by $\mathcal{M}_i^{(b)}$. Similarly, the quantity $\xi_t^{i,j}(\rho, \sigma, \gamma)$ is the responsibility between a transition from state ρ to state σ (reached at time t) for the HMM $\mathcal{M}_j^{(r)}$ and state γ (at time t) of the HMM $\mathcal{M}_i^{(b)}$, when modeling a sequence generated by $\mathcal{M}_i^{(b)}$. Consequently, the statistic $\hat{\nu}_1^{i,j}(\sigma)$ is the expected number of times that the HMM $\mathcal{M}_j^{(r)}$ starts from state σ , when modeling sequences generated by $\mathcal{M}_i^{(b)}$. The quantity $\hat{\nu}^{i,j}(\sigma, \gamma)$ is the expected number of times that the HMM $\mathcal{M}_i^{(b)}$ is state γ when the HMM $\mathcal{M}_j^{(r)}$ is in state σ , when both modeling sequences generated by $\mathcal{M}_i^{(b)}$. Finally, the quantity $\hat{\xi}^{i,j}(\rho, \sigma)$ is the expected number of transitions from state ρ to state σ of the HMM $\mathcal{M}_j^{(r)}$, when modeling sequences generated by $\mathcal{M}_i^{(b)}$.

3.5 M-step

The M-steps involves maximizing (28) with respect to $\mathcal{M}^{(r)}$ for fixed \mathbf{z} , ϕ and η , i.e.,

$$\hat{\mathcal{M}}^{(r)} = \operatorname{argmax}_{\mathcal{M}^{(r)}} \mathcal{J}(\mathcal{M}^{(r)}, \mathbf{z}, \phi, \eta). \quad (44)$$

In the following, we detail the update rules for the parameters of the reduced model $\mathcal{M}^{(r)}$.

3.5.1 HMMs mixture weights

The re-estimation of the mixture weights, given the constraint $\sum_{j=1}^{K^{(r)}} \omega_j^{(r)} = 1$, is solved using the result in Section (6.1):

$$\omega_j^{(r)*} = \frac{\sum_{i=1}^{K^{(b)}} \hat{z}_{i,j}}{K^{(b)}}. \quad (45)$$

3.5.2 Initial state probabilities

The cost function (28) factors independently for each $\{\pi_\sigma^{(r),j}\}_{\sigma=1}^N$ (j is fixed) and reduces to terms in the form:

$$\mathcal{J}(\mathcal{M}^{(r)}, \mathbf{z}, \phi, \eta) = \sum_{\sigma=1}^N \sum_{i=1}^{K^{(b)}} \hat{z}_{i,j} N_i \hat{\nu}_1^{i,j}(\sigma) \log \pi_\sigma^{(r),j}. \quad (46)$$

Considering the constraint $\sum_{\sigma=1}^N \pi_\sigma^{(r),j} = 1$, the results in Section 6.1 gives the update formulas

$$\pi_\sigma^{(r),j*} = \frac{\sum_{i=1}^{K^{(b)}} \hat{z}_{i,j} N_i \sum_{\gamma=1}^N \hat{\nu}_1^{i,j}(\sigma)}{\sum_{\sigma'=1}^N \sum_{i=1}^{K^{(b)}} \hat{z}_{i,j} N_i \sum_{\gamma=1}^N \hat{\nu}_1^{i,j}(\sigma')}. \quad (47)$$

3.5.3 State transition probabilities

Similarly, the cost function (28) factors independently for each $\{a_{\rho,\sigma}^{(r),j}\}_{\sigma=1}^N$ (j and ρ are fixed) and reduces to terms in the form:

$$\mathcal{J}(\{a_{\rho,\sigma}^{(r),j}\}_{\sigma=1}^N, \mathbf{z}, \phi, \eta) = \sum_{\sigma=1}^N \sum_{i=1}^{K^{(b)}} \hat{z}_{i,j} N_i \hat{\xi}^{i,j}(\rho, \sigma) \log a_{\rho,\sigma}^{(r),j} \quad (48)$$

Considering the constraint $\sum_{\sigma=1}^N a_{\rho,\sigma}^{(r),j} = 1$, the results in Section 6.1 gives the update formula

$$a_{\rho,\sigma}^{(r),j*} = \frac{\sum_{i=1}^{K^{(b)}} \hat{z}_{i,j} N_i \hat{\xi}^{i,j}(\rho, \sigma)}{\sum_{\sigma'=1}^N \sum_{i=1}^{K^{(b)}} \hat{z}_{i,j} N_i \sum_{t=2}^T \hat{\xi}^{i,j}(\rho, \sigma')}. \quad (49)$$

3.5.4 Emission probability density functions

In general, in the cost function (28) factors independently for each j, ρ, ℓ , and reduces to terms in the form:

$$\begin{aligned} \mathcal{J}(\mathcal{M}_{\rho,\ell}^{(r),j}, \mathbf{z}, \phi, \eta) = \\ \sum_{i=1}^{K^{(b)}} \hat{z}_{i,j} N_i \sum_{\gamma=1}^N \hat{\nu}^{i,j}(\rho, \gamma) \sum_{m=1}^M c_{\beta,m}^{(b),i} \hat{\eta}_{\ell|m}^{(i,\beta),(j,\rho)} \left(\log c_{\rho,\ell}^{(r),j} + L_G(\mathcal{M}_{\beta,m}^{(b),i} || \mathcal{M}_{\rho,\ell}^{(r),j}) \right) \end{aligned} \quad (50)$$

Using basic matrix calculus, and defining a weighted sum operator

$$\Omega_{j,\rho,\ell}(x(i, \beta, m)) = \sum_i \hat{z}_{i,j} N_i \sum_{\beta} \hat{\nu}_t^{i,j}(\rho, \beta) \sum_{m=1}^M c_{\beta,m}^{(b),i} x(i, \beta, m) \quad (51)$$

the parameters $\mathcal{M}^{(r)}$ are updated accordingly to:

$$c_{\rho,\ell}^{(r),j*} = \frac{\Omega_{j,\rho,\ell} \left(\hat{\eta}_{\ell|m}^{(i,\beta),(j,\rho)} \right)}{\sum_{\ell'=1}^M \Omega_{j,\rho,\ell'} \left(\hat{\eta}_{\ell'|m}^{(i,\beta),(j,\rho)} \right)} \quad (52)$$

$$\mu_{\rho,\ell}^{(r),j*} = \frac{\Omega_{j,\rho,\ell} \left(\eta_{\ell|m}^{(i,\beta),(j,\rho)} \mu_{\beta,m}^{(b,i)} \right)}{\Omega_{j,\rho,\ell} \left(\hat{\eta}_{\ell|m}^{(i,\beta),(j,\rho)} \right)} \quad (53)$$

$$\Sigma_{\rho,\ell}^{(r),j*} = \frac{\Omega_{j,\rho,\ell} \left(\hat{\eta}_{\ell|m}^{(i,\beta),(j,\rho)} \right) \left[\Sigma_{\beta,m}^{(b,i)} + (\mu_{\beta,m}^{(b,i)} - \mu_{\rho,\ell}^{(r),j})(\mu_{\beta,m}^{(b,i)} - \mu_{\rho,\ell}^{(r),j})^t \right]}{\Omega_{j,\rho,\ell} \left(\hat{\eta}_{\ell|m}^{(i,\beta),(j,\rho)} \right)} \quad (54)$$

Equations (52-54) are all weighted averages over all base models, model states, and Gaussian components.

4 Experiments

In this section, we present three empirical studies of the VHEM-H3M algorithm. Each application exploits some of the benefits of VHEM. First of all, instead of clustering HMMs on the parameter manifold, VHEM-H3M clusters HMMs directly through the distributions they represent. Given a collection of input HMMs, VHEM estimates a smaller mixture of novel HMMs that consistently models the distribution represented by the input HMMs. This is achieved by maximizing the log-likelihood of “virtual” samples generated from the input HMMs. As a result, the VHEM cluster centers are consistent with the underlying generative probabilistic framework.

Second, VHEM allows to estimate models from large-scale data sets, by breaking the learning problem into smaller pieces. First, a data set is split into small non-overlapping portions and intermediate models are learned from each portion. Then, the final model is estimated from the intermediate models using the VHEM-H3M algorithm. While based on the same maximum-likelihood principles as direct EM estimation on the full data set, this VHEM estimation procedure has significantly lower memory requirements, since it is no longer required to store the entire data set during parameter estimation. In addition, since the intermediate models are estimated independently of each other, this estimation task can easily be parallelized. Lastly, the “virtual” samples (i.e., sequences) VHEM implicitly generates for maximum-likelihood estimation need not be of the same length as the actual input data for estimating the intermediate models. Making the virtual sequences relatively short will positively impact the run time of each VHEM iteration. This may be achieved without loss of modeling accuracy, as VHEM allows to compensate for shorter virtual training sequences by implicitly integrating over a virtually unlimited number of them.

4.1 Hierarchical motion clustering

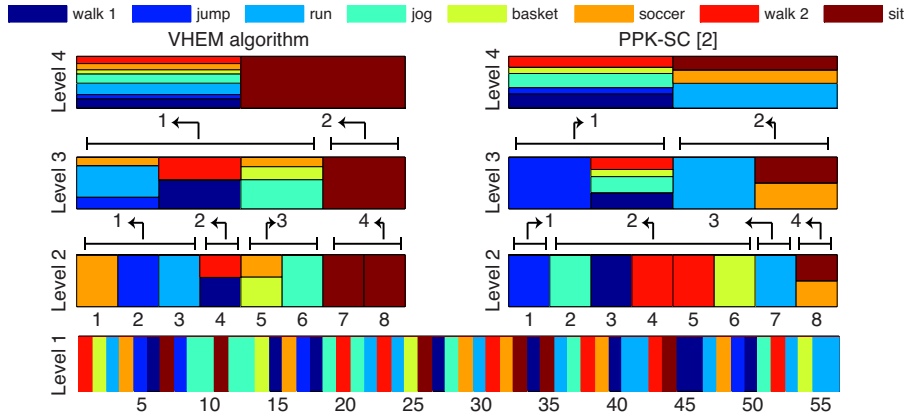


Figure 1: Hierarchical clustering of the MoCap dataset, with VHEM and SC-PPK.

In this experiment we tested the VHEM algorithm on hierarchical motion clustering, using the Motion Capture dataset (<http://mocap.cs.cmu.edu/>), which is a collection of time-series data representing human motions. In particular, we start from $K_1 = 56$ motion examples from 8 different classes, and learn a HMM for each of them, forming the first level of the hierarchy. A tree-structure is formed by successively clustering HMMs with the VHEM algorithm, and using the learned cluster centers as the representative HMMs at the new level. The second, third and fourth levels of the hierarchy correspond to, respectively, $K_2 = 8$, $K_3 = 4$ and $K_4 = 2$.

The hierarchical clustering obtained with VHEM is illustrated in Figure 1 (left). In the first level, each vertical bar represents a motion sequence, with different colors indicating different ground-truth classes. In the second level, the 8 HMM clusters are shown with vertical bars, with the colors indicating the proportions of the motion classes in the cluster. Almost all clusters are populated by examples from a single motion class (e.g., “run”, “jog”, “jump”), which demonstrates that VHEM can group similar motions together. We note an error of the VHEM in clustering a portion of the “soccer” examples with “basket”. Moving up the hierarchy, the VHEM algorithm clusters similar motions classes together (as indicated by the arrows), and at the last (Level 4) it creates a dichotomy between the “sit” and the rest of the motion classes. This is a desirable behavior as the kinetics of the “sit” sequences (i.e., sitting on a stool and going down) are considerably different from the rest. On the right of Figure 1, the same experiment is repeated using spectral clustering in tandem with PPK similarity (SC-PPK) [5]. The SC-PPK clusters motions sequences properly, however it incorrectly aggregates the “sit” and “soccer”, and produces a last level (Level 4) not well interpretable.

While VHEM has lower Rand-index than SC-PPK at Level 2 (0.940 vs. 0.973), it has higher Rand-index at Level 3 (0.775 vs. 0.737) and Level 4 (0.591 vs. 0.568). This suggests that the novel HMMs cluster centers learned by VHEM retain more information than the spectral cluster centers.

	annotation			MAP	retrieval			classification	
	P	R	F		AROC	P@10		VHEM-H3M	EM-H3M
HEM-H3M	0.470	0.210	0.258	0.438	0.700	0.450	$\tau = 5$	0.569	0.349
EM-H3M	0.415	0.214	0.248	0.423	0.704	0.422	$\tau = 10$	0.570	0.389
HEM-DTM	0.430	0.202	0.252	0.439	0.701	0.453	$\tau = 15$	0.573	0.343

Table 1: Annotation and retrieval performance on CAL500, for VHEM-H3M, EM-H3M and HEM-DTM[13]

Table 2: Online hand-writing classification accuracy (20 characters)

4.2 Automatic music tagging

In this experiment we evaluated VHEM-H3M on automatic music tagging. We considered the CAL500 collection from Barrington et al. [12], which consists in 502 songs and provides binary annotations with respect to a vocabulary \mathcal{V} of 149 tags, ranging from genre and instrumentation, to mood and usage. To represent the acoustic content of a song we extract a time series of audio features $\mathcal{Y} = \{y_1, \dots, y_T\}$, by computing the first 13 Mel frequency cepstral coefficients (MFCC) [1] over half-overlapping windows of 92ms of audio signal, augmented with first and second derivatives.

Automatic music tagging is formulated as a supervised multi-class labeling problem [11], where each class is a tag from \mathcal{V} . We model tags with H3M probability distributions over the space of audio fragments (e.g., sequences of $\tau = 125$ audio features, which approximately corresponds to 6 seconds of audio). Each tag model is learned from audio-fragments extracted from relevant songs in the database, using the VHEM-H3M. The database is first processed at the song level, using the EM algorithm to learn a H3M for each song from a dense sampling of audio fragments. For each tag, the song-level H3Ms that are relevant to the tag are pooled together to form a big H3M, and the VHEM algorithm is used to learn the final tag-model.

In table 1 we present a comparison of the VHEM-H3M algorithm with the standard EM-H3M algorithm and a state-of-the-art auto-tagger (HEM-DTM) [13], which uses the dynamic texture mixture model and an efficient HEM algorithm, on both annotation and retrieval on the CAL500 dataset. Annotation is measured with precision (P), recall (R), f-score (F), and retrieval is measured with mean average precision (MAP), area under the operating characteristic curve (AROC), and precision at the first 10 retrieved objects (P@10). All reported metrics are averages over the 98 tags that have at least 30 examples in CAL500, and are result of 5 fold-cross validation. VHEM-H3M achieves better

performance over EM-H3M (except on precision and AROC which are comparable) and strongly improves the top of the ranking list, as demonstrated by the higher P@10 score. Performance of VHEM-H3M and HEM-DTM are close on all metrics with only slight variations, except on annotation precision where VHEM-H3M registers a significantly higher score.

4.3 Online hand-writing recognition

In this experiment we investigated the performance of the VHEM-H3M algorithm on classification of on-line hand-writing. We considered the Character Trajectories Data Set [19], which consists in 2858 examples of characters from the same writer, and used half of the data for training and half for testing. An HMM (with $N = 2$ and $M = 1$) was first learned from each of the training sequences using the EM algorithm. For each letter, all the relevant HMMs were clustered with the VHEM to form a H3M with $K^{(r)} = 2$ components. We repeated the same experiment using the EM-H3M algorithm directly on all the relevant sequences in the train data. For each letter, we allowed the EM algorithm up to three times the total running time of the VHEM (including the estimation of the corresponding intermediate HMMs). Table 2 lists classification accuracy on the test set, for VHEM-H3M, using different values of τ , and for the corresponding runs of EM-H3M. A small τ suffices to provide a regular estimate, and simultaneously determines shorter running times for VHEM (under 2 minutes for all 20 letters). On the other hand, the EM algorithm needs to evaluate the likelihood of all the original sequences at each iteration, which determines slower iterations, and prevents the EM from converging to effective estimates in the time allowed.

5 Conclusion

In this paper, we present a variational HEM (VHEM) algorithm for clustering HMMs through their distributions. Moreover, VHEM summarizes each cluster by estimating a new HMM as cluster center. We demonstrate the efficacy of this algorithm for various applications, including hierarchical motion clustering, semantic music annotation, and online hand-writing recognition.

6 Appendix on useful optimization problems

6.1

The optimization problem

$$\begin{aligned} \max_{\alpha_\ell} \quad & \sum_{\ell=1}^L \beta_\ell \log \alpha_\ell \\ \text{s.t.} \quad & \sum_{\ell=1}^L \alpha_\ell = 1 \\ & \alpha_\ell \geq 0, \forall \ell \end{aligned} \tag{55}$$

is optimized by

$$\alpha_\ell^* = \frac{\beta_\ell}{\sum_{\ell'=1}^L \beta_{\ell'}}. \tag{56}$$

This can be easily computed with the optimization

$$\{\alpha_\ell^*\} = \underset{\alpha_\ell}{\operatorname{argmax}} \sum_{\ell=1}^L \beta_\ell \log \alpha_\ell + \lambda \left(\sum_{\ell=1}^L \alpha_\ell - 1 \right)$$

where the second term is a Lagrangian term for the weights to sum to 1, and noticing that the positivity constraints are automatically satisfied by (56).

6.2

The optimization problem

$$\begin{aligned} \max_{\alpha_\ell} \quad & \sum_{\ell=1}^L \alpha_\ell (\beta_\ell - \log \alpha_\ell) \\ \text{s.t.} \quad & \sum_{\ell=1}^L \alpha_\ell = 1 \\ & \alpha_\ell \geq 0, \forall \ell \end{aligned} \tag{57}$$

is optimized by

$$\alpha_\ell^* = \frac{\exp \beta_\ell}{\sum_{\ell'=1}^L \exp \beta_{\ell'}}. \tag{58}$$

This can be easily computed with the optimization

$$\{\alpha_\ell^*\} = \underset{\alpha_\ell}{\operatorname{argmax}} \sum_{\ell=1}^L \alpha_\ell (\beta_\ell - \log \alpha_\ell) + \lambda \left(\sum_{\ell=1}^L \alpha_\ell - 1 \right)$$

where the second term is a Lagrangian term for the weights to sum to 1, and noticing that the positivity constraints are automatically satisfied by (58).

References

- [1] L. Rabiner and B. H. Juang, *Fundamentals of Speech Recognition*. Upper Saddle River (NJ, USA): Prentice Hall, 1993.
- [2] Y. Qi, J. Paisley, and L. Carin, “Music analysis using hidden markov mixture models,” *Signal Processing, IEEE Transactions on*, vol. 55, no. 11, pp. 5209–5224, 2007.
- [3] R. Nag, K. Wong, and F. Fallside, “Script recognition using hidden markov models,” in *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP’86.*, vol. 11. IEEE, 1986, pp. 2071–2074.
- [4] A. Krogh, M. Brown, I. Mian, K. Sjolander, and D. Haussler, “Hidden markov models in computational biology. applications to protein modeling,” *Journal of Molecular Biology*, vol. 235, no. 5, pp. 1501–1531, 1994.
- [5] T. Jebara, Y. Song, and K. Thadani, “Spectral clustering and embedding with hidden markov models,” *Machine Learning: ECML 2007*, pp. 164–175, 2007.
- [6] T. Jebara, R. Kondor, and A. Howard, “Probability product kernels,” *The Journal of Machine Learning Research*, vol. 5, pp. 819–844, 2004.
- [7] N. Vasconcelos and A. Lippman, “Learning mixture hierarchies,” in *Advances in Neural Information Processing Systems*, 1998.
- [8] A. B. Chan, E. Coviello, and G. Lanckriet, “Derivation of the hierarchical EM algorithm for dynamic textures,” City University of Hong Kong, Tech. Rep., 2010.
- [9] N. Vasconcelos, “Image indexing with mixture hierarchies,” in *IEEE Conf. Computer Vision and Pattern Recognition*, 2001.
- [10] A. Chan, E. Coviello, and G. Lanckriet, “Clustering dynamic textures with the hierarchical em algorithm,” in *Intl. Conference on Computer Vision and Pattern Recognition*, 2010.
- [11] G. Carneiro, A. Chan, P. Moreno, and N. Vasconcelos, “Supervised learning of semantic classes for image annotation and retrieval,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 3, pp. 394–410, 2007.
- [12] D. Turnbull, L. Barrington, D. Torres, and G. Lanckriet, “Semantic annotation and retrieval of music and sound effects,” *IEEE Transactions on Audio, Speech and Language Processing*, vol. 16, no. 2, pp. 467–476, February 2008.
- [13] E. Coviello, L. Barrington, A. Chan, and G. Lanckriet, “Automatic music tagging with time series models,” in *Proceedings ISMIR*, 2010.

- [14] J. Hershey, P. Olsen, and S. Rennie, “Variational Kullback-Leibler divergence for hidden Markov models,” in *Automatic Speech Recognition & Understanding, 2007. ASRU. IEEE Workshop on.* IEEE, 2008, pp. 323–328.
- [15] P. Smyth, “Clustering sequences with hidden markov models,” in *Advances in neural information processing systems*, 1997.
- [16] M. Jordan, Z. Ghahramani, T. Jaakkola, and L. Saul, “An introduction to variational methods for graphical models,” *Machine learning*, vol. 37, no. 2, pp. 183–233, 1999.
- [17] T. S. Jaakkola, “Tutorial on Variational Approximation Methods,” in *In Advanced Mean Field Methods: Theory and Practice.* MIT Press, 2000, pp. 129–159.
- [18] J. Hershey and P. Olsen, “Approximating the Kullback Leibler divergence between Gaussian mixture models,” in *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, vol. 4. Ieee, 2007.
- [19] B. Williams, M. Toussaint, and A. Storkey, “Extracting motion primitives from natural handwriting data,” *Artificial Neural Networks–ICANN 2006*, pp. 634–643, 2006.